9-1-2010

Identifying Finite Mixtures in Econometric Models

Marc Henry

Yuichi Kitamura

Bernard Salanié

IDENTIFYING FINITE MIXTURES IN ECONOMETRIC MODELS

By

Marc Henry, Yuichi Kitamura and Bernard Salanié

September 2010
Revised January 2013

COWLES FOUNDATION DISCUSSION PAPER NO. 1767

COWLES FOUNDATION FOR RESEARCH IN ECONOMICS
YALE UNIVERSITY
Box 208281
New Haven, Connecticut 06520-8281

http://cowles.econ.yale.edu/

Partial Identification of Finite Mixtures in Econometric Models[1]

Marc Henry[2], Yuichi Kitamura[3] and Bernard Salanié[4]

January 7, 2013

[1]Parts of this paper were written while Henry was visiting the University of Tokyo Graduate School of Economics and Salanié was visiting the Toulouse School of Economics; they both thank their respective hosts, the CIRJE and the Georges Meyer endowment for their support. Support from the NSF (grants SES-0551271, SES-0851759, and SES-1156266) and the SSHRC (grant 410-2010-242) is also gratefully acknowledged. The authors thank Koen Jochmans, Ismael Mourifié, Elie Tamer and three anonymous referees for their very helpful comments and suggestions.

[2]Département de sciences économiques, Université de Montréal, CIRANO, CIREQ. E-mail: marc.henry@umontreal.ca.

[3]Economics Department, Yale University. E-mail: yuichi.kitamura@yale.edu.

[4]Economics Department, Columbia University. E-mail: bs2237@columbia.edu.

Abstract

We consider partial identification of finite mixture models in the presence of an observable source of variation in the mixture weights that leaves component distributions unchanged, as is the case in large classes of econometric models. We first show that when the number J of component distributions is known a priori, the family of mixture models compatible with the data is a subset of a $J(J-1)$-dimensional space. When the outcome variable is continuous, this subset is defined by linear constraints which we characterize exactly. Our identifying assumption has testable implications which we spell out for $J = 2$. We also extend our results to the case when the analyst does not know the true number of component distributions, and to models with discrete outcomes.

Introduction

Finite mixture models feature prominently in many areas of econometrics. When individual heterogeneity in labor markets is characterized by a finite number of types, as in Eckstein and Wolpin (1990) and Keane and Wolpin (1997), structural parameters of interest are recovered from a finite mixture. In measurement error models including data contamination and misclassification of treatment or other observed discrete regressors (see Chen, Hong, and Nekipelov (2011)) observed outcomes are drawn from a finite mixture of distributions. The very large class of dynamic models with hidden discrete state variables, such as regime switching, also falls in the category of finite mixtures (see Kim and Nelson (1999) for an extensive treatment). Finally, social interactions, imperfect competition or complementarities in discrete choice models often generate multiple equilibria, hence finite mixture models, where the components are outcome distributions conditional on a realized equilibrium and the equilibrium selection mechanism characterizes the mixture weights.

The statistical literature on parametric estimation of finite mixtures and determination of the number of components in mixtures is vast, as evidenced in a recent account by Frühwirth-Schnatter (2006). Recently, however, attention was drawn to the empirical content of structural economic models with unobserved types or states short of parametric assumptions on component distributions and mixture weights. Several strategies for the nonparametric identification of finite mixtures have emerged as a result. Mahajan (2006), Lewbel (2007) and Hu (2008) rely on instrumental variables to identify models with misclassified discrete regressors. Chen, Hong, and Tamer (2005) rely on auxiliary data and Chen, Hu, and Lewbel (2008, 2009) use shape and moment restrictions to identify several types of measurement error models. Kitamura (2003) relies on shape invariance to identify finite mixture models nonparametrically.

An exclusion restriction, namely a variable shifting mixture weights without affecting component distributions, is maintained in many of the studies mentioned above (namely Mahajan (2006), Lewbel (2007), Hu (2008) and Chen, Hu, and Lewbel (2008, 2009). This exclusion restriction has much larger appeal than the data combination and misclassification framework. It can be derived from the widely maintained Markov assumption in regime-switching and other hidden state models. We also show how it can be substantiated in models of unobserved heterogeneity, where geographical variables, for instance, may shift type proportions without affecting utility; and in models with multiple equilibria, where specific interventions may increase the likelihood of one equilibrium being selected without affecting outcomes conditional on equilibrium.

This exclusion restriction is generically insufficient for nonparametric identification of the component distributions and the mixture weights. However, it has a non-trivial empirical content, which we characterize through a constructive description of the identified set.

Ours is not the first attempt at partial identification of mixture models. Some of the recent work on partial identification studied particular mixture models and/or identifying restrictions. Thus, Horowitz and Manski (1995) derive sharp bounds on the distribution of contaminated variables; but they assume an upper bound on the probability of contamination, while we do not restrict mixture weights. Bollinger (1996) derives sharp bounds on $\mathbb{E}[Y|X]$ when X is a mismeasured binary regressor; our results apply to regressors of any form in any kind of mixture. Hall and Zhou (2003) studied nonparametric identification in models with repeated measurements. More precisely, they derive bounds for the distribution of a T-dimensional mixture when $T \geq 2$ and each component has independent marginals. Kasahara and Shimotsu (2009) build on similar ideas to identify finite mictures of persistent types in dynamic discrete choice models. Bonhomme, Jochmans, and Robin (2012) show point identification when $T \geq 3$ under a rank condition, and they propose a convenient estimation method.

Molinari (2008) gives general partial identification results for the distribution of a misclassified categorical variable. She proposes a direct misclassification approach to the treatment of data errors, which fully exploits all known restrictions on the matrix of data misclassification probabilities. In the model $P_w = P_{w|x} P_x$, Molinari derives sharp bounds on the vector of true frequencies P_x based on the distribution of misclassified data P_w and a very comprehensive class of restrictions on the matrix of misclassification probabilities $P_{w|x}$. In contrast to Molinari (2008), we consider unrestricted outcome variables (continuous and discrete) and we rely on an exclusion restriction rather than on assumptions on the misclassification process.

In the case of a two-component mixture, we show that the identified set can be characterized as a two-parameter family of component distributions and mixture weights. Going beyond the two-component mixture case, we characterize the identified set for a J-component mixture as a $J(J-1)$-parameter family. The extension bears resemblance to Cross and Manski (2002) (and Molinari and Peski (2006)), especially as in both cases the construction requires computation of the extreme points of a convex polytope. But the problem Cross and Manski (2002) study is "ecological inference": the mixture weights are known.

Based on our constructive characterization of the identified set, we provide strategies for the construction of confidence regions. Our bounds are sharp; and our identifying restriction implies testable implications, which are quite simple for $J = 2$ at least.

In general, misspecification of finite mixture models in the form of an erroneous maintained number of component distributions is a serious concern, as it may invalidate inference. This is one of the major themes in the statistical literature on parametric mixtures, since any misspecification of the components may bias the estimate of their number. In econometrics, some recent papers have therefore taken up testing for the true number of components (in Kasahara and Shimotsu (2011) for instance). Our partial identification analysis removes this concern: we show that it can be embedded in a iterative procedure that determines the true number of components without resorting to any parametric assumption.

The paper is organized as follows. Section 1 presents the analytical framework and discusses the exclusion restriction that underlies our partial identification results. To convey the intuition, we first study in section 2 mixtures with two components; Section 3 then gives general results in the J-component case. These two sections mainly focus on continuously distributed outcomes; Section 4 extends our results to discrete outcomes. We also present in Section 4 a iterative procedure to determine the number of components when it is not known a priori. Most proofs are in Appendix B.

1 Finite mixtures with exclusion restrictions

1.1 Analytical framework

Let Y be a random variable and $Z = (X, W)$ a random vector defined on the same probability space. In all that follows, F will denote conditional cumulative distribution functions and lower case letters w, x, y, z will be used to denote realizations of the random elements W, X, Y and Z. We assume that observed outcomes Y are generated from a finite mixture of at most $J \geq 1$ component distributions:

Assumption 1 (Mixture) *For almost all y, z,*

$$F(y|z) = \sum_{j=0}^{J-1} \lambda_j(z) F_j(y|z) \tag{1.1}$$

where the $\lambda_j(z)$ are non-negative numbers and the $F_j(\cdot|z)$ are cumulative distribution functions.

Note that since we assume that both F and the F_j's are cdfs, (1.1) implies that $\sum_{j=0}^{J-1} \lambda_j(z) \equiv 1$. In particular, the non-negativity of the weights implies that none of them can be larger than 1. On the other hand, we allow for the possibility that some of them are actually zero, so that the model has fewer than J components for some or all values of z.

We assume that an infinite sample from the distribution of (Y, Z) is available, so that we can recover the distribution function $F(y|z)$ of Y conditional on Z. The objects of interest are the latent component distributions $F_j(y|z)$ and the mixture weights $\lambda_j(z)$ for $j = 0, \ldots, J-1$. Without further assumptions, the components of the mixture are clearly not identifiable; the observed distribution function $F(y|z)$ could be rationalized as $F(y|z) = \sum_{j=0}^{J-1} \lambda_j(z) F_j(y|z)$ with $\lambda_j = 1$ for $j = 0$, say, and zero otherwise.

The identifying restriction we consider is a source W of variation in the mixture weights that leaves each of the component distributions unchanged. Our whole analysis is conditional on X; and our identification results apply for any value of x for which:

Assumption 2 (Exclusion restriction) *$F_j(y|x,w) = F_j(y|x)$, for all $j = 0,\ldots,J-1$ and all (y,w) in the support of $(Y,W)|X$.*

Let x be one such value. For simplicity, we shall drop x from the notation from now on; all quantities considered will implicitly be functions of x.

We shall be concerned in this paper with the characterization of the empirical content of Assumptions 1 and 2 above. This will take the form of a constructive characterization of the identified set, which we now define:

Definition 1 (Identified Set) *The identified set is the set of distributions $F_j(y|x)$ and mixture weights $z \mapsto \lambda_j(z)$, $j = 0,\ldots,J-1$, that satisfy Assumptions 1 and 2.*

Under Assumptions 1 and 2, the mixture can be written as follows for any pair w, w' in the support of W.

$$\begin{aligned} F(y|w) &= F(y|w') + \sum_{j=0}^{J-1}\big(\lambda_j(w) - \lambda_j(w')\big)\; F_j(y) \\ &= F(y|w') + \sum_{j=1}^{J-1}\big(\lambda_j(w) - \lambda_j(w')\big)\; \big(F_j(y) - F_0(y)\big), \end{aligned}$$

where the first equation results from the exclusion restriction and the second equation results from the mixture specification with $\lambda_0(w) = 1 - \sum_{j=1}^{J-1}\lambda_j(w)$ for all w. Hence the observable $F(y|w) - F(y|w')$ is a $J-1$ dimensional scalar product. The first term

$$\Big(F_j(y) - F_0(y)\Big)_{j=1}^{J-1}$$

is a function of y only. The second term

$$\Big(\lambda_j(w) - \lambda_j(w')\Big)_{j=1}^{J-1}$$

is an additively separable, antisymmetric function of w and w' only. This decomposition will be key to our partial identification results; it will also allow us to construct overidentification tests of Assumptions 1 and 2.

1.2 Discussion of the exclusion restriction

The variables in w function as traditional (nonparametric) instruments: under Assumption 2 they can vary without changing the distribution of any individual component. In the case of regime switching models, Assumption 2 is a direct implication of the usual Markov assumption. In misclassification and data contamination models, it is equivalent to the mismeasured variable being non-informative on the outcome, conditional on the true value of the variable. This is again a common (but not universal) assumption. In structural economic models with discrete unobserved heterogeneity or with multiple equilibria, the validity of Assumption 2 depends on the context. In each case, we also need the component weights λ_j to depend on w "enough" that they give the instruments identifying power. We now discuss these four applications in more detail. As we will see later, while Assumption 2 may be more or less convincing in a given application, it is a testable assumption.

1.2.1 Regime switching and Markov decision models

Consider the classical Markov switching model (see Kim and Nelson (1999) for a survey), where Y_t is independently and identically distributed conditionally on a state variable S_t that follows a Markov chain. For simplicity, let S_t be binary: $S_t \in \{0, 1\}$, with transition probabilities

$$\Pr(S_t = 1|S_{t-1} = 1) = P_{11} \quad \text{and} \quad \Pr(S_t = 0|S_{t-1} = 0) = P_{00}.$$

Assumptions 1 and 2 are automatically satisfied in this model, with $Y = Y_t$ and $W = Y_{t-1}$. Indeed, denoting $\lambda(w) \equiv \lambda_1(w) \equiv \Pr(S_t = 1|Y_{t-1} = w)$, we have

$$F(y_t|Y_{t-1} = w) = \lambda(w)F(y_t|S_t = 1) + (1 - \lambda(w))F(y_t|S_t = 0).$$

Moreover, it is easy to see that $\lambda(w)$ does depend on w, unless Y_t is independent of S_t and/or $P_{11} = P_{00}$. Special cases include mean switching, with y_t i.i.d. conditionally on S_t and $\mu_{S_t} = S_t\mu_1 + (1-S_t)\mu_2$, and stochastic volatility, with y_t i.i.d. conditionally on $Var(y_t) = \sigma^2_{S_t} = S_t\sigma^2_1 + (1 - S_t)\sigma^2_2$.

This example can easily be extended to m-dependence: if there exists an $m \geq 1$ with

$$F(y_t|S_t = s, y_{t-1}, \dots, y_1) \equiv F(y_t|S_t = s, y_{t-1}, \dots, y_{t-m}),$$

for all states s and all $t \geq m$, then the variable $Z = (Y_{t-1}, \ldots, Y_1)$ can be split into $X = (Y_{t-1}, \ldots, Y_{t-m})$ and $W = (Y_{t-m-1}, \ldots, Y_1)$. In particular, Assumption 2 holds in any model in which the observed trajectory is a finite-order autoregression conditionally on the hidden Markov chain.

1.2.2 Data contamination and misclassification

Models with measurement error on a discrete regressor also often satisfy Assumption 2. Consider an outcome Y affected by an unobserved treatment T^* taking values $T^* = t_0, t_1, \ldots, t_{J-1}$. Let T be an observed variable that is informative on $T^\star$. T could be misclassified treatment, as in Aigner (1973) for the binary case and Molinari (2008) for any discrete regressor; more generally, it could be any measurement that is correlated with T^*. As before, additional conditioning variables X could be incorporated without any substantive change.

The classical assumption on misclassification error, as imposed in most of the recent literature on misclassified treatment surveyed in Chen, Hong, and Nekipelov (2011), posits independence of classification error and outcome conditionally on the true treatment:

$$Y \perp\!\!\!\perp T \,|\, T^*.$$

Then the conditional distribution function of outcome Y conditional on measurement T is

$$F_{Y|T}(y|T) = \sum_{j=0}^{J-1} F_{Y|T,T^\star}(y|T^* = t_j, T) \Pr(T^* = t_j|T) = \sum_{j=0}^{J-1} F_{Y|T^\star}(y|T^* = t_j) \Pr(T^* = t_j|T);$$

and it satisfies Assumptions 1 and 2 with $W = T$, $\lambda_j(w) = \Pr(T^* = t_j|T = w)$ and $F_j(y) = F_{Y|T^\star}(y|T^* = t_j)$. The weights λ_j depend on w in so far as the measurement T is informative on the true treatment T^*.

The classical assumption on misclassification error comes at a cost; we would not expect it to hold when misclassification error is correlated to non-compliance, or the extent of misreporting depends on unobservable individual heterogeneity. The Assumption 2 would not apply in general.

1.2.3 Models with unobserved heterogeneity

Consider a general structural microeconometric model, where observed outcomes Y are functions $Y = g(S, Z, \varepsilon)$ of observed heterogeneity $Z = (W, X)$, a discrete unobservable agent type $S = s_0, s_1, \ldots, s_{J-1}$ and an error term ε. Then Assumption 2 holds when

$$y \perp\!\!\!\perp W \mid S, X;$$

for instance when $g(S, Z, \varepsilon) \equiv g(S, W, \varepsilon)$ and ε is independent of (S, Z). The instruments W will be a source of identifying power if the distribution of S depends on W as well as on X.

If Y represents the demand for a good for instance, we require the unobserved heterogeneity in demand to be adequately summarized by the combination of an agent type S and an idiosyncratic shock ε; and we can use any instrument W that does not enter preferences or covariates X, and yet changes the distribution of agent types. Geographical variables fit the bill as long as they do not directly enter utilities. In dynamic settings such as Markov decision processes, we can also appeal to past observations as in Section 1.2.1. Finally, variables that are not in buyers' information sets at the time of purchase satisfy the first criterion, and the second one too if they change the composition of demand. We develop in Appendix A a simple oligopoly model to illustrate this last point.

1.2.4 Multiple equilibria

Economic models of imperfect competition, social interactions and joint investment with spillovers typically incorporate non-cooperative games in which multiple equilibria are the norm rather than the exception. With a finite set of equilibria $E = e_0, \ldots, e_{J-1}$, realized outcomes are generated as a mixture[1]:

$$F_{Y|Z}(y|z) = \sum_{j=0}^{J-1} \Pr(E = e_j | Z = z) F_{Y|E,Z}(y|e_j, z).$$

Assumption 2 then holds if a variable W does not affect outcomes conditional on the realized equilibrium:

$$Y \perp\!\!\!\perp W \mid E, X.$$

[1]We thank Elie Tamer for pointing out this class of applications of the mixture model.

For identification, we also need W to enter the equilibrium selection mechanism $\Pr(E = e_j|Z = z)$. We now discuss several frameworks in which Assumption 2 is reasonable.

Policy interventions that affect the equilibrium selection are prime candidates as instruments W. In the oligopolistic competition analysis of Ciliberto and Tamer (2009), policies aimed at reducing collusion among firms may affect equilibrium selection differentially in regional markets. There is also a sizable literature in macroeconomics and development on coordination failures. In their theory of the Big Push, Murphy, Shleifer, and Vishny (1989) propose subsidizing fixed entry costs in joint investments with spillovers to prevent poverty traps. More generally, fixed cost shifters that do not affect pricing conditional on entry are potential instruments in problems of joint investment with spillovers (see e.g. Hendricks and Kovenock (1989) for information spillovers.) Cooper and Corbae (2002) explain financial collapse through coordination failure in market participation. In this framework, Ennis and Keister (2006) argue that lower tax rates are likely to increase the probability of the Pareto efficient equilibrium being selected; but other types of intervention, such as subsidies, are more likely to be outcome neutral conditional on equilibrium and hence satisfy Assumption 2. In Forbes and Rigobon (2002), financial contagion is defined as a jump from a low correlation equilibrium to a high correlation equilibrium. Similarly, Pesaran and Pick (2007) argue that policy interventions are more likely to be effective if "the cause of a crisis is a random jump between equilibria, i.e., contagion" than if "a crisis spreads to other markets because the fundamentals are correlated." This is exactly the spirit of our Assumption 2.

When social interactions are prevalent, the regional heterogeneity of outcomes across time and space is often attributed to multiple equilibria. The *tipping point* theory of segregation in Schelling (1971) is an early example. The model of wage discrimination through negative stereotypes of Coate and Loury (1993) and the model of criminal activities of Calvó-Armengol and Zenou (2004) also exhibit such multiple equilibria. In all of these cases, history dependence and variations in social norms induce variation in the equilibrium selection mechanism; but they are typically excluded from utility and hence leave outcomes conditional on equilibrium unchanged. Any such source of variation can serve as an identifying W.

2 Partial Identification of two-component mixtures

From now on, we shall maintain Assumptions 1 and 2 throughout and characterize their empirical content with a constructive characterization of the identified set of Definition 1. Start with the case where the mixture is known to involve exactly two component distributions. We will denote $\lambda_1(w)$ simply by $\lambda(w)$, and $\lambda_0(w) = 1-\lambda(w)$.

As discussed in Section 1, to complement the exclusion restriction of Assumption 2, we need minimal variation in the mixture weights. Also, the existence of exactly two components implies restrictions. We posit

Assumption 3

(i) $\Pr(0 < \lambda(W)) \Pr(\lambda(W) < 1) > 0$ *and* $\Pr(F_0(Y) = F_1(Y)) < 1$.

(ii) There exist w_0 *and* w_1 *in the support of* W *such that* $\lambda(w_0) \neq \lambda(w_1)$.

Assumption 3.(i) implies that the mixture does not degenerate to one component; and (ii) will ensure that w has identifying power. Note that under Assumptions 1 and 2, Assumption 3 could only fail if $F(\cdot|w)$ were independent of w, which is clearly testable.

Since

$$F(y|w_1) - F(y|w_0) = (\lambda(w_1) - \lambda(w_0)) (F_1(y) - F_0(y)),$$

the left-hand side is non-zero at any y where F_0 and F_1 do not coincide. At any such y, for any w we have

$$\frac{F(y|w) - F(y|w_0)}{F(y|w_1) - F(y|w_0)} = \frac{\lambda(w) - \lambda(w_0)}{\lambda(w_1) - \lambda(w_0)}. \tag{2.1}$$

Therefore the left-hand-side of this equation is a function of w only, which we denote $\Lambda(w)$. It is identified from the data, and by construction $\Lambda(w_0) = 0$ and $\Lambda(w_1) = 1$.

From (2.1), we obtain a two-parameter characterization of the mixture weights that are compatible with the data:

$$\lambda(w) = \phi + \psi\Lambda(w), \tag{2.2}$$

where $\phi = \lambda(w_0)$ and $\psi = \lambda(w_1) - \lambda(w_0)$. Once the parameters ϕ and ψ are fixed, the component distributions are also identified. Defining $\delta = F_1 - F_0$, we have

$$\delta(y) = F_1(y) - F_0(y) = \frac{F(y|w_1) - F(y|w_0)}{\lambda(w_1) - \lambda(w_0)} = \frac{1}{\psi}[F(y|w_1) - F(y|w_0)]. \quad (2.3)$$

By construction,

$$\begin{aligned} F_0(y) &= F(y|w_0) - \lambda(w_0)\delta(y), \\ F_1(y) &= \Delta(y) + F_0(y) \\ &= F(y|w_0) + [1 - \lambda(w_0)]\delta(y). \end{aligned}$$

Since, by definition, $\phi = \lambda(w_0)$ and $\psi = \lambda(w_1) - \lambda(w_0)$ and since the latter is non zero by Assumption 3, we obtain the two-parameter family characterization for the component distributions.

$$F_0(y) = F(y|w_0) - \frac{\phi}{\psi}[F(y|w_1) - F(y|w_0)], \quad (2.4)$$

$$F_1(y) = F(y|w_0) + \frac{1-\phi}{\psi}[F(y|w_1) - F(y|w_0)] \quad (2.5)$$

The identified set for the mixture under Assumptions 1-3 is therefore determined by the set of admissible values for the pair (ϕ, ψ). Such a pair is admissible if and only if $\lambda(w)$ is a probability and the two component distributions $F_0(y)$ and $F_1(y)$ are cdfs.

- First consider the constraints on the weight: $0 \leq \lambda(w) \leq 1$ for all w. Defining

  $$\overline{\Lambda} = \sup_w \Lambda(w) \text{ and } \underline{\Lambda} = \inf_w \Lambda(w), \quad (2.6)$$

  these result in two necessary and sufficient conditions on the pair (ϕ, ψ):

  $$0 \leq \phi + \psi\overline{\Lambda} \leq 1 \text{ and } 0 \leq \phi + \psi\underline{\Lambda} \leq 1.$$

  These conditions (which imply $\phi > 0$ but do not restrict the sign of ψ) are equivalent to $-\psi\overline{\Lambda} \leq \phi \leq 1 - \psi\overline{\Lambda}$ and $-\psi\underline{\Lambda} \leq \phi \leq 1 - \psi\underline{\Lambda}$, and finally to

  $$-\min(\psi\overline{\Lambda}, \psi\underline{\Lambda}) \leq \phi \leq 1 - \max(\psi\overline{\Lambda}, \psi\underline{\Lambda}). \quad (2.7)$$

  The inequalities above can be expressed in terms of the reparametrization $(-\phi/\psi, (1-\phi)/\psi)$ as

  $$\min\left(\frac{1-\phi}{\psi}, \frac{-\phi}{\psi}\right) \leq \underline{\Lambda} \leq 0 \leq 1 \leq \overline{\Lambda} \leq \max\left(\frac{1-\phi}{\psi}, \frac{-\phi}{\psi}\right).$$

- Let us proceed to the constraints on the component distributions: F_0 and F_1 should be non-decreasing, right-continuous and with left and right limits 0 and 1. It follows directly from Equations (2.4) and (2.5) that the left and right limits of F_0 and F_1 are 0 and 1, and that they are right-continuous. Now consider the monotonicity constraints. For two realizations $y' > y$ of Y, denote $D_k(y,y') = F(y'|w_k) - F(y|w_k) \geq 0$ for $k = 0, 1$. We must have

$$D_0(y,y') + \zeta\left(D_1(y,y') - D_0(y,y')\right) \geq 0$$

for both $\zeta = -\phi/\psi$ and $\zeta = (1-\phi)/\psi$. This is equivalent to the two conditions

$$\sup_{y'>y;D_1(y,y')>D_0(y,y')} \frac{-D_0(y,y')}{D_1(y,y') - D_0(y,y')} \leq \min\left(-\frac{\phi}{\psi}, \frac{1-\phi}{\psi}\right)$$

and

$$\max\left(-\frac{\phi}{\psi}, \frac{1-\phi}{\psi}\right) \leq \inf_{y'>y;D_1(y,y')<D_0(y,y')} \frac{D_0(y,y')}{D_0(y,y') - D_1(y,y')}.$$

These two conditions, along with (2.7), give the sharp bounds on (ϕ, ψ) and therefore on (λ, F_0, F_1). When outcomes y are continuously distributed, the analysis is simpler since the monotonicity constraints become constraints on the densities.

Assumption 4 *The observable distribution $F(.|w)$ is differentiable for all y in the support of Y and all w in the support of W.*

Under Assumption 4, the monotonicity of F_0 and F_1 is equivalent to the non-negativity of their densities:

$$\begin{aligned} f_* := \sup_{f(y|w_1)>f(y|w_0)} \frac{-f(y|w_0)}{f(y|w_1) - f(y|w_0)} &\leq \min\left(-\frac{\phi}{\psi}, \frac{1-\phi}{\psi}\right) \leq 0 \\ \leq \max\left(-\frac{\phi}{\psi}, \frac{1-\phi}{\psi}\right) &\leq \inf_{f(y|w_0)>f(y|w_1)} \frac{f(y|w_0)}{f(y|w_0) - f(y|w_1)} := f^*. \quad (2.8) \end{aligned}$$

Denote the likelihood ratio

$$r(y) := \frac{f(y|w_1)}{f(y|w_0)}.$$

Since densities have total mass 1,

$$\int (r(y)-1)f(y|w_0)dy = 0$$

and so

$$\underline{r} := \inf_{y\in Y} r(y) < 1 < \sup_{y\in Y} r(y) := \overline{r}.$$

Then

$$f_* = -\frac{1}{\overline{r}-1} \text{ and } f^* = \frac{1}{1-\underline{r}}. \tag{2.9}$$

We therefore have the following characterization of the identified set in the case of two-component mixtures with continuous outcomes—we treat the case of discrete outcomes y separately in section 4.2.

Theorem 1 (Two-component identified set with continuous outcomes) *Under Assumptions 1, 2, 3 and 4, the component mixtures and mixture weights are identified as a two parameter family according to (2.2), (2.4), and (2.5); and the identified set for the parameter pair (ϕ,ψ) is*

$$\{(\phi,\psi) : f_* \leq \min((1-\phi)/\psi, -\phi/\psi) \leq \underline{\Lambda} \text{ and } \overline{\Lambda} \leq \max((1-\phi)/\psi, -\phi/\psi) \leq f^*\},$$

where the identified parameters $\underline{\Lambda}$ and $\overline{\Lambda}$ are defined in (2.6) and f_* and f^* in (2.8).

The bounds can be equivalently written in terms of (ψ,ϕ) as

$$\max(-\underline{\Lambda}\psi, -\overline{\Lambda}\psi, \min(1-\psi f_*, 1-\psi f^*)) \leq \phi \leq \min(1-\underline{\Lambda}\psi, 1-\overline{\Lambda}\psi, \max(-\psi f_*, -\psi f^*)).$$

While these inequalities look complex, note that this is in great part due to the "labeling problem": if we decide for instance to call "component 1" the component whose weight is larger in w_1 than in w_0, then $\psi = \lambda(w_1) - \lambda(w_0) > 0$ and the bounds on (ψ,ϕ) simplify to

$$\max(-\underline{\Lambda}\psi, 1-\psi f^*)) \leq \phi \leq \min(1-\overline{\Lambda}\psi, -\psi f_*).$$

Figures 1 and 2 represent the identified region for the pair (ψ,ϕ) and the corresponding region for $(-\phi/\psi,(1-\phi)/\psi)$, restricted to $\psi > 0$. The identified region with $\psi < 0$ is symmetric with respect to the $\psi = 0$ axis in Figure 1; and it is obtained by a rotation of angle $-\pi/2$ around the origin in Figure 2.

It follows from (2.7) and (2.8) that the projection of the identified set on the ψ axis is a symmetric pair of intervals:

$$(f^* - f_*)^{-1} \leq |\psi| \leq (\overline{\Lambda} - \underline{\Lambda})^{-1}, \tag{2.10}$$

which shows the impact of variation in W and in Y on the size of the identified region. If W induces a large variation in the distribution of Y then, by the definition of $\Lambda(w)$, the bounds $\underline{\Lambda}$ and $\overline{\Lambda}$ will be farther apart and the identified set for (ψ, ϕ) will shrink. Similarly, a large variation in the density of Y conditional on W will pull the bounds f^* and f_* closer together and shrink the identified set. This can be seen from Figures 1 and 2: a larger support for W leads to an increase in $\overline{\Lambda} - \underline{\Lambda}$, and hence to a smaller identification region.

Note that the model is point identified when $f_* = \underline{\Lambda}$ and $f^* = \overline{\Lambda}$; again this is testable. Theorem 1 also shows that the model is rejected when $f_* > \underline{\Lambda}$ or $f^* < \overline{\Lambda}$. This provides a test of specification of the model, which involves testing jointly the exclusion restriction and the hypothesis that there are two component distributions in the mixture. We will build on this idea in section 4.1 when we describe our iterative procedure to determine the number of components J.

Using equation (2.9), it is easy to see that

- the model is rejected iff

$$\underline{r} < 1 - \frac{1}{\overline{\Lambda}} \text{ or } \overline{r} > 1 - \frac{1}{\underline{\Lambda}}$$

- it is point identified if both of these inequalities are replaced with equalities
- and it is partially identified otherwise.

Point-identification may seem like a rare case; but there are useful classes of models for which the two conditions are binding. If for instance the range of the true likelihood ratio $R(y) = f_1(y)/f_0(y)$ includes 0 and $+\infty$, then

$$\underline{r} = \min\left(\frac{\lambda(w_1)}{\lambda(w_0)}, \frac{1-\lambda(w_1)}{1-\lambda(w_0)}\right) \text{ and } \overline{r} = \max\left(\frac{\lambda(w_1)}{\lambda(w_0)}, \frac{1-\lambda(w_1)}{1-\lambda(w_0)}\right)$$

and the model is point-identified whenever $\lambda(w_0) \neq \lambda(w_1)$, as it does under Assumption 3(ii). Additional a priori restrictions, such as a monotone likelihood ratio assumption on $R(y)$, would allow the analyst to relax these conditions.

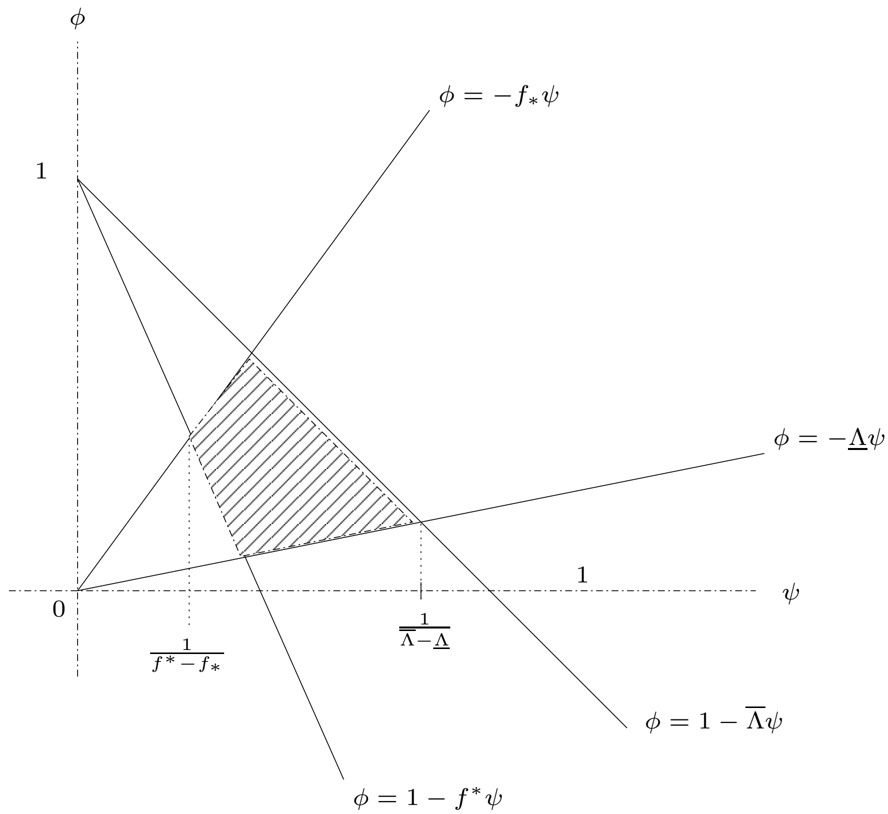

Figure 1: The shaded area is the identified region for the pair (ψ, ϕ) in the half-plane $\psi > 0$.

Note from (2.3) that any linear functional of $(F_1 - F_0)$ is identified up to scale. Denote E_i the expectation operator with respect to F_i. Then for any function h of y, we can test whether $E_1 h(Y) - E_0 h(Y)$ is zero simply by testing that $E\left(h(Y)|W = w\right)$ depends on w. If it does, then for any other function g of y, the ratio

$$\frac{E_1 g(Y) - E_0 g(Y)}{E_1 h(Y) - E_0 h(Y)}$$

is point-identified.

In the context of a model with randomized assignment and mismeasured treatment, this ratio is simply a relative average treatment effect. Take h to be the identity

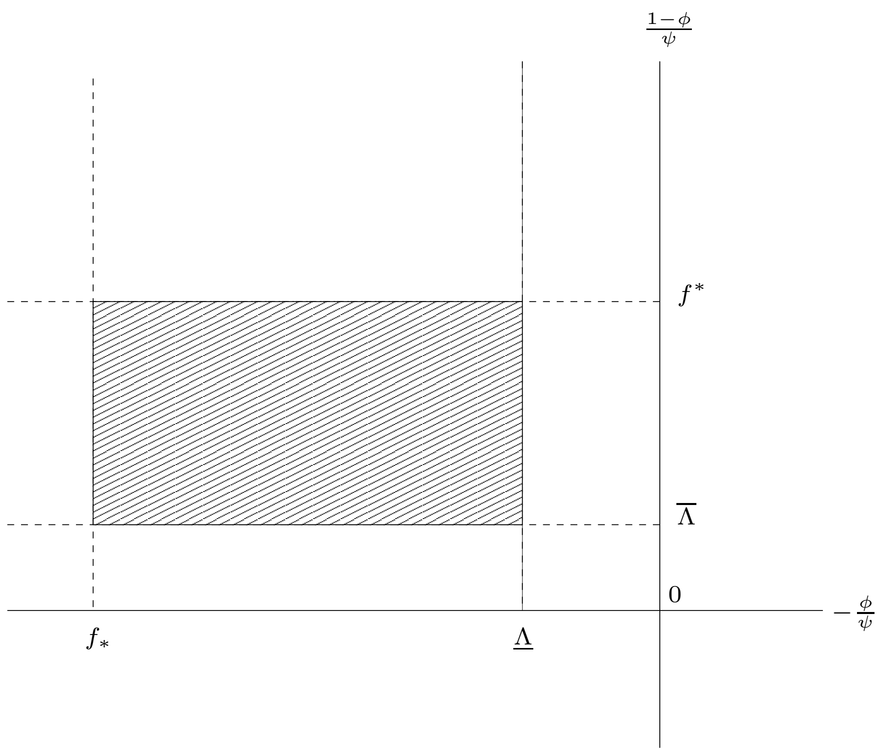

Figure 2: The shaded area is the identified region for the pair $(-\phi/\psi, (1-\phi)/\psi)$ which parameterize the two component distributions F_1 and F_0, restricted to $\psi > 0$.

function for instance, while $g(y) = \mathbb{1}(y \geq a)$. Then if the average treatment effect on Y is non-zero, the relative quantile treatment effects

$$\frac{\mathrm{Pr}_1(Y \geq a) - \mathrm{Pr}_0(Y \geq a)}{E_1 Y - E_0 Y}$$

are point-identified for all values of a.

3 Finite mixtures of arbitrary order

We now turn to general partial identification results. We first assume that the true number of mixture components is known and equal to J. The next section will extend the identification results to the case of an unknown number of mixture components. Under Assumptions 1 and 2, we recall that for any (y, w, w_0),

$$F(y|w) - F(y|w_0) = \sum_{j=1}^{J-1} (\lambda_j(w) - \lambda_j(w_0))(F_j(y) - F_0(y)) = \boldsymbol{\psi}(w)^t\ \boldsymbol{\delta}(y),$$

where (dropping the dependence on w_0 from the notation) we define $\boldsymbol{\psi}(w)$ as the $(J-1)$-vector with j-th component $\psi_j(w) := \lambda_j(w) - \lambda_j(w_0)$ and $\boldsymbol{\delta}(y)$ as the $(J-1)$-vector with j-th component $\delta_j(y) := F_j(y) - F_0(y)$. As in the case of two components, we need sufficient variability of mixture weights to complement the exclusion restriction of Assumption 2. We therefore state the analogue of Assumption 3 in the case of J component distributions:

Assumption 5 (Relevance) *There exist $(w_0, w_1, \ldots, w_{J-1})$ in the support of W such that the $(J-1) \times (J-1)$ matrix $\mathbf{\Psi}$ with j-th column $\boldsymbol{\psi}(w_j)$ is invertible.*

Note that Assumption 5 immediately implies an order condition: the support of W must contain at least J distinct points. Under Assumption 5, let $\mathbf{h_c}(y)$ denote the $(J-1)$-vector with j-th component $F(y|w_j) - F(y|w_0)$. Then

$$\mathbf{h_c}(y) = \mathbf{\Psi}^t\ \boldsymbol{\delta}(y),$$

so that $\boldsymbol{\delta}(y) = \left(\mathbf{\Psi}^t\right)^{-1} \mathbf{h_c}(y)$. This translates immediately into the identification of component distributions as a $J(J-1)$ parameter family:

$$\text{for all } j = 0, \ldots, J-1,\ F_j(y) = F(y|w_0) + (\mathbf{e_j} - \boldsymbol{\phi})^t \left(\mathbf{\Psi}^t\right)^{-1}\ \mathbf{h_c}(y), \tag{3.1}$$

where $\mathbf{e_j}$ is the unit vector with a 1 in the j-th row, with the convention that $\mathbf{e_0}$ is the zero vector; and $[\;]_j$ denotes the j-th component of the vector inside the brackets. The component distributions are identified in Equation (3.1) up to the $J(J-1)$ unknown parameters that define $\boldsymbol{\phi}$ and $\boldsymbol{\Psi}$, since all other quantities involved, namely $F(y|w_0)$ and $\mathbf{h_c}(y)$, are point-identified.

Now assume that there is sufficient variation in $\boldsymbol{\delta}(y)$:

Assumption 6 (Rank) *There exist $(y_1, \ldots, y_{J-1})$ in the support of Y such that the $(J-1) \times (J-1)$ matrix $\boldsymbol{\Delta}$ with j-th column $\boldsymbol{\delta}(y_j)$ is invertible.*

Again, an order condition immediately arises: under Assumption 6, Y must have at least J distinct points of support. Note that if the number of distinct component distributions is assumed to be exactly equal to J, this order condition is automatically satisfied.

Assumptions 5 and 6 both relate to unobservable quantities. We could alternatively have used Assumption 7, which is directly testable from the data:

Assumption 7 *There exist $(w_0, \ldots, w_{J-1})$ in the support of W and $(y_1, \ldots, y_{J-1})$ in the support of Y such that the $(J-1) \times (J-1)$ matrix $\mathbf{H}$ with j-th column $\mathbf{h_c}(y_j)$ is invertible.*

The $(J-1) \times (J-1)$ matrix $\mathbf{H}$ is the product of the two $(J-1) \times (J-1)$ matrices $\boldsymbol{\Psi}$ and $\boldsymbol{\Delta}$. The following lemma follows immediately:

Lemma 1 (Testability of rank conditions) *Under Assumptions 1 and 2, Assumptions 5 and 6 are equivalent to Assumption 7.*

Under Assumptions 5 and 6 (or Assumption 7), we can now identify the mixture weights as a $J(J-1)$ family. Indeed, for all y, w, we have:

$$\begin{aligned} F(y|w) - F(y|w_0) &= \boldsymbol{\psi}(w)^t \, \boldsymbol{\delta}(y) \\ &= \boldsymbol{\psi}(w)^t \, \left(\boldsymbol{\Psi}^t\right)^{-1} \, \mathbf{h_c}(y), \end{aligned}$$

so that, denoting $\mathbf{h_r}(w)$ the identified $(J-1)$-vector with j-th component $F(y_j|w) - F(y_j|w_0)$, we have

$$\mathbf{h_r}(w)^t = \boldsymbol{\psi}(w)^t \left(\boldsymbol{\Psi}^t\right)^{-1} \mathbf{H}$$

and we finally obtain identification of the mixture weights as a two-parameter family.

More precisely, call $\boldsymbol{\lambda}(w)$ the (unknown) vector of mixture weights with j-th component $\lambda_j(w)$:

$$\boldsymbol{\lambda}(w) = \boldsymbol{\phi} + \boldsymbol{\psi}(w) = \boldsymbol{\phi} + \boldsymbol{\Psi} \left(\mathbf{H}^t\right)^{-1} \mathbf{h_r}(w), \tag{3.2}$$

where $\boldsymbol{\Lambda}(w) = \left(\mathbf{H}^t\right)^{-1} \mathbf{h_r}(w)$ is the analogue of the identified $\Lambda(w)$ function of the two-component case. In order to characterize the identified set, we only need to derive sharp bounds for $(\boldsymbol{\phi}, \boldsymbol{\Psi})$. As in the case of the two-component mixture, we obtain these bounds by imposing probability constraints on $\boldsymbol{\lambda}(w)$ and monotonicity constraints on the component distributions $F_j(y)$, $j = 0, 1, \ldots, J-1$:

- *Probability constraints:* we need $\mathbf{0} \leq \boldsymbol{\lambda}(w)$ and $\mathbf{1}^t\boldsymbol{\lambda}(w) \leq 1$ on the mixture weights. Hence we require

  $$\mathbf{0} \leq \boldsymbol{\phi} + \boldsymbol{\Psi} \left(\mathbf{H}^t\right)^{-1} \mathbf{h_r}(w) \text{ and } \mathbf{1}^t \left(\boldsymbol{\phi} + \Psi \left(\boldsymbol{H}^t\right)^{-1} \mathbf{h_r}(w)\right) < 1$$

  for all w in the support of W. These are linear inequalities in (Φ, Ψ); as such, they only need to be imposed at the extreme points of convex hull of the range of $w \mapsto \boldsymbol{\Lambda}(w) = \left(\mathbf{H}^t\right)^{-1} \mathbf{h_r}(w)$.

- *Monotonicity constraints:* As with the case of two components, equation (3.1) implies directly that the F_j's range from 0 to 1. We will again treat the case of discrete supports separately; here we assume that the cdfs of outcomes are differentiable, as in Assumption 4. Denote $f(y|w)$ the density of outcomes conditional on w and $\mathbf{h}'_\mathbf{c}(y)$ the derivative of $\mathbf{h_c}(y)$; the monotonicity constraints on the component distributions are

  $$\text{for all } j = 0, 1, \ldots, J-1, \; f(y|w_0) + (\mathbf{e_j} - \boldsymbol{\phi})^t \left(\boldsymbol{\Psi}^t\right)^{-1} \mathbf{h}'_\mathbf{c}(y) \geq 0$$

  for all y in the domain of Y. These inequalities are not linear in $(\boldsymbol{\phi}, \boldsymbol{\Psi})$ any more; but they are linear in the transformed parameters $\boldsymbol{\Omega_j} = (\mathbf{e_j} - \boldsymbol{\phi})^t \left(\boldsymbol{\Psi}^t\right)^{-1}$. Therefore they only need to be checked at the extreme points of the range of the function $\boldsymbol{\mathcal{F}}(y) := -\mathbf{h}'_\mathbf{c}(y)/f(y|w_0)$.

The previous discussion is summarized in the following theorem, which we prove in Appendix B:

Theorem 2 (Identified set) *The identified set for the component distributions and the mixture weights under Assumptions 1, 2, 4 and 7 is the $J(J-1)$ parameter family defined by Equations (3.1)-(3.2) along with the following constraints on $(\boldsymbol{\phi}, \boldsymbol{\Psi})$:*

- *the linear constraints $\boldsymbol{\phi} + \boldsymbol{\Psi}\mathbf{e} > 0$ and $\mathbf{1}^t(\boldsymbol{\phi} + \Psi\mathbf{e}) < 1$ for all extreme points $\mathbf{e}$ of the convex hull of the range of the identified function $w \mapsto \boldsymbol{\Lambda}(w) = (\mathbf{H}^t)^{-1}\mathbf{h}_{\mathbf{r}}(w)$;*
- *the quadratic constraints $\mathbf{f}^t\boldsymbol{\Psi}^{-1}(\mathbf{e_j} - \boldsymbol{\phi}) \leq 1$ for $j = 0, \ldots, J-1$ and for all extreme points $\mathbf{f}$ of the convex hull of the range of the identified function $y \mapsto \boldsymbol{\mathcal{F}}(y) := -\mathbf{h}'_{\mathbf{c}}(y)/f(y|w_0)$.*

The hypotheses of Theorem 2 preclude discrete outcomes and require a priori knowledge of the true number of component distributions. The next section shows that these limitations are superficial, as the same reasoning can be applied to discrete outcomes and unknown mixture order. Section 4.2 also shown how to considerably reduce the computational burden associated with the construction of the identified set, with a view to form confidence regions with traditional partial identification inference procedures.

4 Extensions

We now move beyond the assumptions of Theorem 2 to consider the determination of the order J of the mixture, and the case of discrete-valued outcomes.

4.1 Determining J

Theorem 2 assumed that the analyst knows the exact number of distinct component distributions. In fact, a simple iterative procedure allows us to determine the number of components and the identified set for the component distributions and mixture weights[2].

Start with $J = 2$. Note that the true number of components is at least 2 under assumption 3.

[2]We thank Ismael Mourifié for suggesting this iterative procedure to us.

1. Construct the identified set according to the procedure of section 3.
2. If Step 1 yields a nonempty identified set, the mixture model with at most J components cannot be rejected. The true number of components is then identified as J.
3. If the identified set in Step 1 is empty, the mixture model with a maximum of J components is rejected. Then make $J \leftarrow J+1$ and return to Step 1.

4.2 Discrete outcomes: latent class analysis

The identification results of Theorem 2 rely on Assumption 4, which rules out discrete outcomes. However, most of the analysis carries over with simple changes in notation. To emphasize it, we shall retain the same notation for slightly different objects, with probability mass functions replacing probability distribution functions.

One substantial difference is that the true number of mixture component distributions is directly identified from the matrix of conditional probabilities. Let the support of Y be $\{y_1, \ldots, y_N\}$, and that of W be $\{w_0, w_1, \ldots, w_{M-1}\}$. Write $\Pr(y|w)$ for the probability $\Pr(Y = y|W = w)$ of outcome y conditional on $W = w$. First note that under Assumption 1, there cannot be more than $\min(M, N)$ components. The following lemma identifies the true number of component distributions:

Lemma 2 (Mixture order for finite outcomes) *Under Assumptions 1 and 2 with* $J = \min(M, N)$*, the number of non-zero weights and distinct components is* $(K+1)$*, where* K *is the rank of the* $(M-1)\times N$ *matrix with* (i,j)*-th entry* $\Pr(y_j|w_i)-\Pr(y_j|w_0)$.

With K defined as in Lemma 2 above and suitable relabeling of the supports of Y and W, we can assume that the $K \times K$ matrix $\mathbf{H}$ with (i,j)-th entry $\Pr(y_j|w_i) - P(y_j|w_0)$ is invertible. As before, call $\mathbf{h_r}(w_i)$ its i-th row and $\mathbf{h_c}(y_j)$ its j-th column. Then, following the same reasoning as in Section 3, we obtain identification of the component probabilities P_j and mixture weights λ_j, $j = 0, 1, \ldots, J-1$, as a $J(J-1)$ parameter family, with the $(J-1)$-vector $\boldsymbol{\phi}$ and the $(J-1)\times(J-1)$ matrix $\boldsymbol{\Psi}$ as parameters:

$$P_j(y_l) = P(y_l|w_0) + (\mathbf{e_j} - \boldsymbol{\phi})^t \left(\boldsymbol{\Psi}^t\right)^{-1} \mathbf{h_c}(y_l), \tag{4.1}$$

$$\lambda(w_k) = \boldsymbol{\phi} + \boldsymbol{\Psi} \left(\mathbf{H}^t\right)^{-1} \mathbf{h_r}(w_k) \tag{4.2}$$

for all $l = 1, \ldots, N$, all $k = 1, \ldots, M$ and all $j = 0, 1, \ldots, J-1$, where for $j \geq 1$, $\mathbf{e_j}$ is the unit vector with a 1 in the j-th row, and $\mathbf{e_0}$ is a $(J-1)$ vector of zeros.

Characterizing the identified set for the mixture now requires identifying sharp bounds for the parameter pair $(\boldsymbol{\phi}, \boldsymbol{\Psi})$, which are, as before, $(\mathbf{e_j} - \boldsymbol{\phi})^t \left(\boldsymbol{\Psi}^t\right)^{-1} \mathbf{e} \leq 1$ for $\mathbf{e}$ in the union of the ranges of $y \mapsto -\mathbf{h_c}(y)/P(y|w_0)$ and $y \mapsto \mathbf{h_c}(y)/(1 - P(y|w_0))$; and $0 \leq \boldsymbol{\phi} + \boldsymbol{\Psi}\mathbf{e}$ and $\mathbf{1}^t(\boldsymbol{\phi} + \Psi\mathbf{e}) < 1$ for all extreme points $\mathbf{e}$ of the convex hull of the range of the identified function $w \mapsto \boldsymbol{\Lambda}(w) = \left(\mathbf{H}^t\right)^{-1} \mathbf{h_r}(w)$.

Consider now the computational aspects of the problem of checking whether a particular choice of $(\boldsymbol{\phi}, \boldsymbol{\Psi})$ belongs to the identified set, hence whether a particular choice of mixture model is admissible. Call A the convex hull of the collection of points in $\mathbb{R}^M$ with coordinates $\mathbf{h_c}(y_l)/P(y_l|w_0)$ or $\mathbf{h_c}(y_l)/(1 - P(y_l|w_0))$, $l = 1, \ldots, N$ and B the convex hull of the collection of points in $\mathbb{R}^N$ with coordinates $\Lambda(w_k)$, $k = 1, \ldots, M$. Checking that a $(\boldsymbol{\phi}, \boldsymbol{\Psi})$ pair is admissible is equivalent to checking the linear constraints $(\mathbf{e_j} - \boldsymbol{\phi})^t \left(\boldsymbol{\Psi}^t\right)^{-1} \mathbf{e} \leq 1$ for all extreme points $\mathbf{e}$ of A and the linear constraints $\boldsymbol{\phi} + \boldsymbol{\Psi}\mathbf{e} > 0$ for all extreme points $\mathbf{e}$ of B. The problem of finding the extreme points of the convex hull of a collection of points is a classical one and numerous algorithms exist (see for instance Matoušek (2002)) for which off-the-shelf implementations abound. The Matlab *ConvexHull* command is one of them. The advantage of the extreme points method are both computational and statistical. First, the linear constraints are checked on a reduced number of points, producing computational efficiency gains. Second and more importantly, it reduces the number of inequalities to check in the construction of a confidence region for the identified set, thereby reducing the conservativeness of the region as in Chernozhukov, Hong, and Tamer (2007).

Concluding Remarks

Finite mixtures are pervasive in econometrics, and yet most of the literature has imposed strong parametric restrictions in order to estimate them. We fully characterized the identified region under an exclusion restriction that is quite natural in some important classes of models.

In the two-component case, point-identification can be obtained under two additional restrictions. One can for instance impose that one component dominates in the left tail and the other one dominates in the right tail. In parallel work, we explore the asymptotic properties of an estimator that relies on tail dominance.

Although the case of two-component mixtures is very important in applications, inference for partially identified finite mixtures of more than two components is a natural next step in this research program. We are currently working to adapt the literature on estimation of partially identified models defined by moment inequalities. Finally, one could combine our exclusion restriction with others in order to achieve tighter identification. The repeated measurement literature is a case in point: the results of Bonhomme, Jochmans, and Robin (2012) for instance can be integrated with ours.

Appendices

A Oligopoly model

Consider an oligopoly with N firms. Each firm i operates with costs of production $C_i(\cdot)$ and faces demand $D_i(p_i, p_{-i}, s)$, where the demand parameter s can take on two values $\overline{s} > \underline{s}$.

The timing of the game and the information structure are the following:

1. Cost functions C_i are realized.
2. Each firm observes its own cost along with a private signal s_i that is informative on other firm's costs and on the state s.
3. Firms simultaneously choose prices p_i to maximize their expected profits.
4. Then s is realized and sales are made.
5. The econometrician later observes noisy measurements of costs, prices, sales, and profits of all firms, which we collect in four N-vectors $\tilde{\mathbf{D}}$, $\tilde{\mathbf{p}}$, $\tilde{\mathbf{C}}$, and $\tilde{\boldsymbol{\pi}}$.

We focus on the distribution of observed sales conditional on observed profits, prices, and costs:

$$\begin{aligned} F(\tilde{\mathbf{D}}|\tilde{\boldsymbol{\pi}}, \tilde{\mathbf{p}}, \tilde{\mathbf{C}}) = {} & F(\tilde{\mathbf{D}}|\tilde{\boldsymbol{\pi}}, \tilde{\mathbf{p}}, \tilde{\mathbf{C}}, s = \overline{s}) \Pr(s = \overline{s}|\tilde{\boldsymbol{\pi}}, \tilde{\mathbf{p}}, \tilde{\mathbf{C}}) \\ & + F(\tilde{\mathbf{D}}|\tilde{\boldsymbol{\pi}}, \tilde{\mathbf{p}}, \tilde{\mathbf{C}}, s = \underline{s}) \Pr(s = \underline{s}|\tilde{\boldsymbol{\pi}}, \tilde{\mathbf{p}}, \tilde{\mathbf{C}}). \end{aligned}$$

Now assume that

1. Prices are observed by the econometrician without measurement error: $\tilde{\mathbf{p}} = \mathbf{p}$.
2. Observed demand and profits are conditionally independent:

$$\tilde{\mathbf{D}} \perp\!\!\!\perp \tilde{\boldsymbol{\pi}} \,|\, (\mathbf{p}, \tilde{\mathbf{C}}, s).$$

When these conditions hold, observed profits $\tilde{\boldsymbol{\pi}}$ do not appear any more in the conditional distributions $F(\tilde{\mathbf{D}}|\tilde{\boldsymbol{\pi}}, \mathbf{p}, \tilde{\mathbf{C}}, s)$, so that Assumption 2 applies with $y = \tilde{\mathbf{D}}$ as the outcome, $w = \tilde{\boldsymbol{\pi}}$ as the instrument, and with covariates x that contain $(\mathbf{p}, \tilde{\mathbf{C}}, s)$.

A variety of more primitive assumptions imply condition 2 above. If measurement errors are classical (independent of all true values), then condition 2 holds if the measurement errors on demand are independent of those on profits and on costs. Both conditions hold for instance in Hendricks, Pinkse, and Porter (2003), where ex-post information is obtained on the value of oil tracts in wildcat lease contracts.

B Proofs

Proof of Theorem 1: Theorem 1 is a special case of Theorem 2. However, proving it directly simplifies notation and helps gain intuition towards the proof of the more general case.

We already showed in the main text that Assumptions 1 to 4 imply the set of inequalities on the pair $((1-\phi)/\psi, \phi/\psi)$ that appears in Theorem 1. The set of inequalities on (ψ, ϕ) follows immediately. We still need to prove that the implied bounds on (F_0, F_1, λ) do not depend on the choice of w_1 and w_0

To see this, take any choice (w_0^1, w_1^1) of (w_0, w_1), along with any (ϕ^1, ψ^1). The corresponding mixture weights and component functions λ^1, F_0^1, F_1^1 are

$$\begin{aligned}
\lambda^1(w) &= \phi^1 + \psi^1 \frac{\lambda(w) - \lambda(w_0^1)}{\lambda(w_1^1) - \lambda(w_0^1)} \\
F_0^1(y) &= F(y|w_0^1) - \frac{\phi^1}{\psi^1}\left(F(y|w_1^1) - F(y|w_0^1)\right) \\
F_1^1(y) &= F(y|w_0^1) + \frac{1-\phi^1}{\psi^1}\left(F(y|w_1^1) - F(y|w_0^1)\right).
\end{aligned}$$

The last two equations can also be rewritten as

$$F_0^1(y) = F_0(y) + \left(\lambda(w_0^1) - \frac{\phi^1}{\psi^1}(\lambda(w_1^1) - \lambda(w_0^1))\right)(F_1(y) - F_0(y)) \tag{B.1}$$

$$F_1^1(y) - F_0^1(y) = \frac{\lambda(w_1^1) - \lambda(w_0^1)}{\psi^1}. \tag{B.2}$$

For any other choice (w_0^2, w_1^2), define (ϕ^2, ψ^2) such that the two functions λ^1 and λ^2

coincide. This is always possible: we only need

$$\frac{\psi^2}{\lambda(w_1^2)-\lambda(w_0^2)} = \frac{\psi^1}{\lambda(w_1^1)-\lambda(w_0^1)}$$
$$\phi^2-\lambda(w_0^2)\frac{\psi^2}{\lambda(w_1^2)-\lambda(w_0^2)} = \phi^1-\lambda(w_0^1)\frac{\psi^1}{\lambda(w_1^1)-\lambda(w_0^1)}.$$

Moreover, equation (B.2) shows that with this choice, $F_1^2 - F_0^2 \equiv F_1^1 - F_0^1$; and it easy to check in equation (B.1) that $F_0^2 \equiv F_0^1$.

We still need to check that if (ϕ^1, ψ^1) satisfies the inequalities in the Theorem for (w_0^1, w_1^1), then (ϕ^2, ψ^2) also does for (w_0^2, w_1^2). But since the former set of inequalities are necessary and sufficient conditions for λ^1 to be a probability and for (F_0^1, F_1^1) to be cdfs, and $(\lambda^2, F_0^2, F_1^2)$ coincides with $(\lambda^1, F_0^1, F_1^1)$, this holds by construction.

Proof of Theorem 2: Again, we only need to show that the constraints are not affected by the choice of $w_0, w_1, \ldots, w_{J-1}$ and $y_1, \ldots, y_{J-1}$.

We proceed as with the proof of Theorem 1. Consider any choice $\mathbf{w^1} = (w_0^1, \ldots, w_{J-1}^1)$ and $\mathbf{y^1} = (y_1^1, \ldots, y_{J-1}^1)$ satisfying Assumptions 5 and 6, and any $(\boldsymbol{\phi^1}, \boldsymbol{\Psi^1})$. Then we construct

$$\boldsymbol{\lambda^1}(w) = \boldsymbol{\phi^1} + \boldsymbol{\Psi^1}\left(\left(\mathbf{H^1}\right)^t\right)^{-1}\mathbf{h^1_r}(w) \tag{B.3}$$

$$F_j^1(y) = F(y|w_0^1) + (\mathbf{e_j} - \boldsymbol{\phi^1})^t((\boldsymbol{\Psi^1})^t)^{-1}\mathbf{h^1_c}(y), \tag{B.4}$$

where $\mathbf{h^1_c}(y)$ is the $(J-1)$ vector with j-th component $F(y|w_j^1) - F(y|w_0^1)$ and $\mathbf{h^1_r}(w)$ is the $(J-1)$ vector with j-th component $F(y_j|w) - F(y_j|w_0^1)$, and $\mathbf{H}$ is the matrix with (i,j)-th element $F(y_j|w_i^1) - F(y_j|w_0^1)$.

Now take an alternative choice $(\mathbf{w^2}, \mathbf{y^2})$ and choose $\boldsymbol{\phi^2}$ and $\boldsymbol{\Psi^2}$ so that $\boldsymbol{\lambda^2} \equiv \boldsymbol{\lambda^1}$. Since $[\mathbf{h^1_r}(w)]_j = (\boldsymbol{\delta}(y_j))^t\,(\boldsymbol{\lambda}(w) - \boldsymbol{\lambda}(w_0^1))$, this boils down to

$$\boldsymbol{\Psi^2}\left(\left(\mathbf{H^2}\right)^t\right)^{-1}\left(\boldsymbol{\Delta^2}\right)^t = \boldsymbol{\Psi^1}\left(\left(\mathbf{H^1}\right)^t\right)^{-1}\left(\boldsymbol{\Delta^1}\right)^t \tag{B.5}$$

$$\boldsymbol{\phi^2} - \boldsymbol{\Psi^2}\left(\left(\mathbf{H^2}\right)^t\right)^{-1}\left(\boldsymbol{\Delta^2}\right)^t\boldsymbol{\lambda}(w_0^2) = \boldsymbol{\phi^1} - \boldsymbol{\Psi^1}\left(\left(\mathbf{H^1}\right)^t\right)^{-1}\left(\boldsymbol{\Delta^1}\right)^t\boldsymbol{\lambda}(w_0^1), \tag{B.6}$$

which are the multidimensional analogs of equations (B.1) and (B.2). They clearly have a unique solution in $(\boldsymbol{\phi^2}, \boldsymbol{\Psi^2})$ under our assumptions.

Moreover, equation (B.5) implies

$$((\mathbf{\Psi^1})^t)^{-1} = ((\mathbf{\Psi^2})^t)^{-1}\mathbf{H^2}\left(\mathbf{\Delta^2}\right)^{-1}\mathbf{\Delta^1}\left(\mathbf{H^1}\right)^{-1}$$

so that, using equation (B.4),

$$\begin{aligned} F_j^1(y) - F_0^1(y) &= (\mathbf{e_j})^t((\mathbf{\Psi^1})^t)^{-1}\mathbf{h^1_c}(y) \\ &= (\mathbf{e_j})^t((\mathbf{\Psi^2})^t)^{-1}\mathbf{H^2}\left(\mathbf{\Delta^2}\right)^{-1}\mathbf{\Delta^1}\left(\mathbf{H^1}\right)^{-1}\mathbf{h^1_c}(y). \end{aligned}$$

Now $[\mathbf{h^1_c}(y)]_j = (\boldsymbol{\delta}(y))^t(\boldsymbol{\lambda}(w_j^1) - \boldsymbol{\lambda}(w_0^1))$, so that... Finally, rewriting (B.4) for $j = 0$ as

$$F_0^2(y) - F_0^1(y) = \left(\boldsymbol{\lambda}(w_0^2) - \boldsymbol{\lambda}(w_0^1)\right)^t \boldsymbol{\delta}(y) - \left(\boldsymbol{\phi^2}\right)^t((\mathbf{\Psi^2})^t)^{-1}\mathbf{h^2_c}(y) + \left(\boldsymbol{\phi^1}\right)^t((\mathbf{\Psi^1})^t)^{-1}\mathbf{h^1_c}(y),$$

we get $F_0^2 \equiv F_0^1$.

We conclude as in the proof of Theorem 1 by noting that $(\boldsymbol{\phi}^2, \mathbf{\Psi^2})$ satisfies the constraints in Theorem 2 for $(\mathbf{w^2}, \mathbf{y^2})$ if $(\boldsymbol{\phi}^1, \mathbf{\Psi^1})$ does for $(\mathbf{w^1}, \mathbf{y^1})$. As before, we have two alternative expressions for the same weights and the same component distributions. One of the expressions satisfies the constraints of Theorem 2, hence so must the other by construction.

References

Aigner, D. (1973): "Regression with a binary independent variable subject to errors of observations," *Journal of Econometrics*, 1, 49–60.

Bollinger, C. (1996): "Bounding mean regressions when a binary regressor is mismeasured," *Journal of Econometrics*, 673, 387–399.

Bonhomme, S., K. Jochmans, and J.-M. Robin (2012): "Nonparametric estimation of finite-mixture models," unpublished manuscript.

Calvó-Armengol, A., and Y. Zenou (2004): "Social networks and crime decisions: the role of social structure in facilitating delinquent behavior," *International Economic Review*, 45, 939–958.

Chen, X., H. Hong, and D. Nekipelov (2011): "Nonlinear models of measurement errors," *Journal of Economic Literature*, 49, 901–937.

Chen, X., H. Hong, and E. Tamer (2005): "Measurement Error Models with Auxiliary Data," *Review of Economic Studies*, 72, 343–366.

Chen, X., Y. Hu, and A. Lewbel (2008): "A note on the closed-form identification of regression models with a mismeasured binary regressor," *Statistics and Probability Letters*, 78, 1473–1479.

——— (2009): "Nonparametric identification of regression models containing a misclassified dichotomous regressor without instruments," *Statistica Sinica*, 1, 1–3.

Chernozhukov, V., H. Hong, and E. Tamer (2007): "Estimation and Confidence Regions for Parameter Sets in Econometric Models," *Econometrica*, 75, 1243–1285.

Ciliberto, F., and E. Tamer (2009): "Market structure and multiple equilibria in airline markets," *Econometrica*, 70, 1791–1828.

Coate, S., and G. Loury (1993): "Will affirmative action policies eliminate negative stereotypes," *American Economic Review*, 83, 1220–1240.

Cooper, R., and D. Corbae (2002): “Financial Collapse: A Lesson from the Great Depression,” *Journal of Economic Theory*, 107, 159–190.

Cross, P., and C. Manski (2002): “Regressions, short and long,” *Econometrica*, 70, 357–368.

Eckstein, Z., and K. Wolpin (1990): “Estimating a market equilibrium search model from panel data on individuals,” *Econometrica*, 58, 783–808.

Ennis, H., and T. Keister (2006): “Bank runs and investment decisions revisited,” *Journal of Monetary Economics*, 53, 217–232.

Forbes, K., and R. Rigobon (2002): “No contagion, only interdependence: Measuring stock market comovements,” *Journal of Finance*, 57, 2223–2251.

Frühwirth-Schnatter, S. (2006): *Finite mixture and Markov switching models.* Springer.

Hall, P., and X.-H. Zhou (2003): “Nonparametric identification of component distributions in a multivariate mixture,” *Annals of Statistics*, 31, 201–224.

Hendricks, K., and D. Kovenock (1989): “Asymmetric information, information externalities and efficiency: the case of oil exploration,” *RAND Journal of Economics*, 20, 164–182.

Hendricks, K., J. Pinkse, and R. Porter (2003): “Empirical implications of equilibrium bidding in first-price, symmetric, common value auctions,” *Review of Economic Studies*, 70, 115–145.

Horowitz, J., and C. Manski (1995): “Identification and robustness via contaminated and corrupt data,” *Econometrica*, 63, 281–302.

Hu, Y. (2008): “Identification and estimation of nonlinear models with misclassification error using instrumental variables,” *Journal of Econometrics*, 144, 27–61.

Kasahara, H., and K. Shimotsu (2009): “Nonparametric identification and estimation of finite mixture models of dynamic discrete choice,” *Econometrica*, 77, 135–175.

——— (2011): “Nonparametric identification and estimation of multivariate mixtures,” Queens Economics Department Working Paper.

Keane, M., and K. Wolpin (1997): “The career decisions of young men,” *Journal of Political Economy*, 105, 473–522.

Kim, C.-J., and C. Nelson (1999): *State-Space Models with Regime-Switching: Classical and Gibbs-Sampling Approaches with Applications*. MIT Press.

Kitamura, Y. (2003): “Nonparametric identifiability of finite mixtures,” unpublished manuscript.

Lewbel, A. (2007): “Estimation of average treatment effects with misclassification,” *Econometrica*, 75, 537–551.

Mahajan, A. (2006): “Identification and estimation of regression models with misclassification,” *Econometrica*, 74, 631–665.

Matoušek, J. (2002): *Lectures on discrete geometry*. Springer.

Molinari, F. (2008): “Partial identification of probability distributions with misclassified data,” *Journal of Econometrics*, 144, 81–117.

Molinari, F., and M. Peski (2006): “Generalization of a result on “Regressions, short and long’,” *Econometric Theory*, 22, 159–163.

Murphy, K., A. Shleifer, and R. Vishny (1989): “Industrialization and the big push,” *Journal of Political Economy*, 65, 1003–1026.

Pesaran, H., and A. Pick (2007): “Econometric issues in the analysis of contagion,” *Journal of Economic Dynamics and Control*, 31, 1247–1277.

Schelling, T. (1971): “Dynamic models of segregation,” *Journal of Mathematical Sociology*, 1, 143–186.